Chapter

Perspective Chapter: Artificial Intelligence-assisted Screening Systems in Ophthalmology – Deep Learning Architectures, Clinical Validation and Implementation Strategies

Kalyan Chakravarthy

Abstract

Artificial intelligence (AI) has transformed eye disease screening. Deep learning systems can now automatically detect diseases from retinal images. This chapter examines current AI technologies, including convolutional neural networks, vision transformers, and foundation models. These systems diagnose diabetic retinopathy, glaucoma, age-related macular degeneration, and anterior segment diseases. Clinical studies show a sensitivity of 87–98% and a specificity of 87–98% across large patient datasets. Implementation requires understanding Food and Drug Administration (FDA) and European Union (EU) regulations, healthcare system integration through Health Level Seven (HL7) Fast Healthcare Interoperability Resources (FHIR) standards, and addressing algorithmic fairness to ensure equal performance across all patient groups. New technologies like federated learning allow hospitals to collaborate on AI training while keeping patient data private. Edge computing enables screening in remote areas without internet access. Vision-language models provide clear explanations of diagnoses in natural language. Current challenges include a 5–15% accuracy loss when systems are moved between hospitals and the need for uncertainty quantification. Future directions focus on combining multiple data sources, including retinal images and electronic health records.

Keywords: artificial intelligence, attention mechanisms, clinical validation, convolutional neural networks, deep learning, diabetic retinopathy, edge computing, federated learning, foundation models, glaucoma screening, healthcare implementation, image classification, medical imaging, ophthalmology, retinal disease, transfer learning, vision transformers

1. Introduction

1.1 Global burden of visual impairment

Visual impairment is a major global health problem. It affects 2.2 billion people worldwide, according to the World Health Organization (WHO) [1]. At least 1 billion of these cases are preventable or treatable.

The main causes include uncorrected refractive errors (123.7 million people), cataracts (65.2 million), age-related macular degeneration (AMD; 8.7 million), glaucoma (6.9 million), and diabetic retinopathy (3 million) [2].

Low and middle-income countries are most affected. Approximately 89% of visually impaired people live in these countries. They often lack access to eye care services and screening programs.

Traditional eye screening faces major barriers. The world has only 40,000 ophthalmologists to serve over seven billion people [3], which creates severe shortages, especially in rural areas.

Geographic inequality makes the problem worse. In Sub-Saharan Africa, there is one ophthalmologist per 1,000,000 people. In high-income nations, the ratio is 1:10,000. These gaps lead to delayed diagnoses and preventable vision loss.

Vision loss costs the global economy $411 billion annually in lost productivity [4]. Healthcare costs significantly add to this burden **Figure 1**.

1.2 Evolution of AI in ophthalmology

Artificial intelligence (AI) offers a solution to healthcare delivery challenges. It provides automated diagnostic capabilities. Deep learning systems, trained on large medical image datasets, now match or exceed human expert performance [5, 6].

The Food and Drug Administration (FDA) approved IDx-DR in April 2018, marking a major milestone. IDx-DR (Digital Diagnostics Inc., Coralville, Iowa) was the first AI system authorized for clinical use without a doctor reviewing its decisions [7]. This proved that fully automated screening works, and it led to the rapid development of similar systems for glaucoma, AMD, and other eye diseases.

Recent advances in foundation models and large language models have added new capabilities [8, 9]. These include generating explanations in natural language and adapting to new diagnostic tasks with minimal training data. Systems can now combine visual information from images with patient records. This marks a shift from narrow, single-task models to flexible AI systems that address multiple screening needs.

The IDx-DR approval established two important AI categories:

(1) **Autonomous AI systems** make diagnostic decisions independently, with minimal human oversight. They are approved as standalone Class II/III medical devices. The manufacturer shares clinical responsibility (e.g., IDx-DR with 87% sensitivity and 91% specificity).
(2) **Decision-support AI systems** help clinicians but require a physician to -confirm decisions before taking action. Clinical responsibility remains with the doctor (as is the case with most current systems). This distinction impacts regulatory pathways (FDA De Novo review: 12–18 months), clinical workflows, and legal liability.

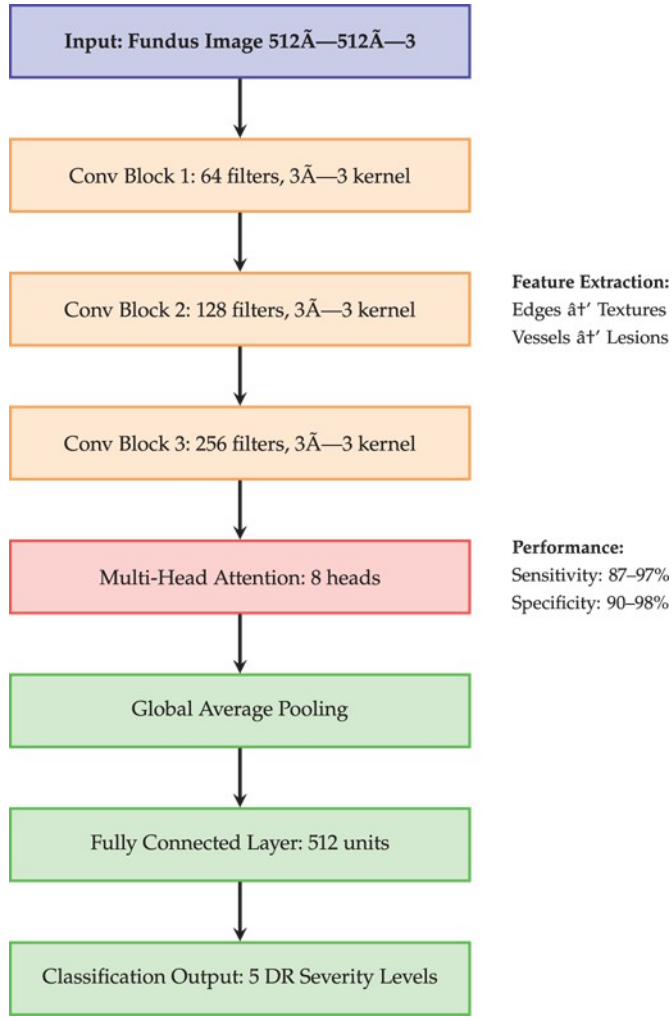
Input: Fundus Image 512Ã—512Ã—3
Conv Block 1: 64 filters, 3Ã—3 kernel
Conv Block 2: 128 filters, 3Ã—3 kernel
Conv Block 3: 256 filters, 3Ã—3 kernel
Multi-Head Attention: 8 heads
Global Average Pooling
Fully Connected Layer: 512 units
Classification Output: 5 DR Severity Levels
Feature Extraction:
Edges â†' Textures
Vessels â†' Lesions
Performance:
Sensitivity: 87–97%
Specificity: 90–98%

Figure 1.
Deep learning CNN architecture for diabetic retinopathy screening, showing vertical information flow through hierarchical layers.

2. Technical foundations of AI screening systems

2.1 Deep learning architectures for retinal image analysis

Modern AI screening systems use deep learning, which has advanced rapidly since 2012 [10]. Convolutional neural networks (CNNs) are the main technology; they analyze retinal images layer by layer. Early layers detect simple features like edges and textures, while deeper layers identify disease signs like microaneurysms, hemorrhages, and exudates [11].

Residual networks (ResNet) improved CNNs in 2016. They solved a technical problem that limited network depth. ResNet can have over 150 layers using "skip connections" that help information flow through the network [12].

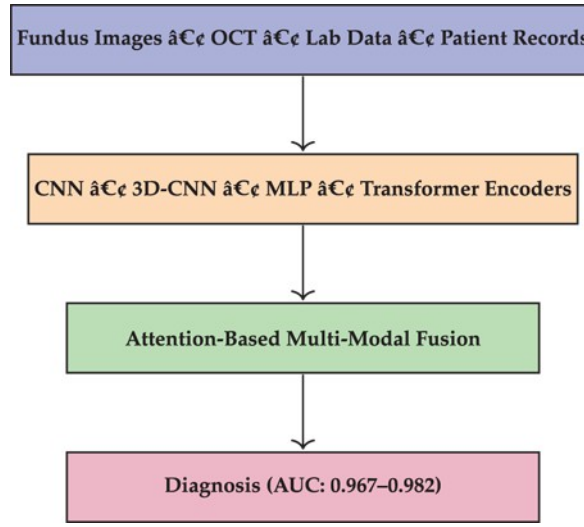
Fundus Images â€¢ OCT â€¢ Lab Data â€¢ Patient Records
CNN â€¢ 3D-CNN â€¢ MLP â€¢ Transformer Encoders
Attention-Based Multi-Modal Fusion
Diagnosis (AUC: 0.967–0.982)

Figure 2.
Multimodal AI integration combines fundus photography, optical coherence tomography (OCT) imaging, laboratory results, and patient records through specialized encoders and attention-based fusion, achieving a diagnostic area under the curve (AUC) of 0.967–0.982.

Vision transformers arrived in 2021, bringing a new approach [13]. Instead of analyzing images section by section like CNNs, they divide images into patches and process them all together. This helps them see relationships across the entire image at once.

Studies show that vision transformers perform 2–5 percentage points better than CNNs for tasks requiring the analysis of distant image regions [14, 15] **Figure 2**.

2.2 Foundation models and transfer learning

Foundation models are a new development in medical AI. They learn from massive datasets and can adapt to new tasks quickly [16]. These models train on hundreds of millions of images and learn general visual patterns. This means they require minimal additional training for specific medical tasks.

The Segment Anything Model (SAM) from Meta AI (2023) demonstrates this capability. It can identify retinal structures like optic disc boundaries, blood vessels, and the fovea without any specific retinal training [17].

Transfer learning has become standard practice [18]. AI models first train on ImageNet, which contains millions of general images across 20,000 categories. This provides a strong foundation for retinal disease classification.

Raghu et al. showed that transfer learning from ImageNet improves diagnostic accuracy by 5–15 percentage points compared to training from random initialization. It also reduces computational requirements by 40–60% through faster convergence [19]. Recent investigations into medical-specific foundation models suggest potential for superior performance. These models are pretrained exclusively on diverse medical imaging modalities, capturing domain-specific visual patterns critical for accurate pathology detection [20].

2.3 Attention mechanisms and interpretability

Attention mechanisms enable neural networks to selectively focus computational resources on diagnostically relevant image regions, simultaneously providing visual explanations for algorithmic decisions [21]. Multihead self-attention is the core component of transformer architectures; it computes pairwise relationships between all spatial locations in retinal images. As a result, models can capture long-range dependencies between distant anatomical structures.

Gradient-weighted Class Activation Mapping (Grad-CAM) generates saliency maps that highlight image regions most influential for classification decisions [22]. It computes the gradients of predicted class scores with respect to the final convolutional layer activations. These visualizations enable clinicians to verify that AI systems base diagnostic decisions on clinically relevant pathological features rather than on spurious correlations or imaging artifacts.

Despite significant progress, interpretability remains a significant challenge in the clinical deployment of AI screening systems. Grad-CAM visualizations often prove too coarse-grained for detailed clinical validation, as they highlight broad anatomical regions rather than specific pathological structures [23]. Recent advances in transformer-based architectures with built-in attention visualization capabilities show promise for improved interpretability. However, extensive clinical validation is required to establish the clinical utility of these explanation methods [24].

2.4 Data preprocessing pipeline

Effective data preprocessing is fundamental to achieving robust AI performance. Image quality directly affects diagnostic accuracy, with poor-quality images reducing performance by 10–15%. A standardized preprocessing pipeline ensures consistent results across different imaging devices and clinical settings **Figure 3**.

Automated quality assessment algorithms evaluate image sharpness, illumination uniformity, and artifact presence. Poor-quality images are flagged for retaking. Studies show that rejecting 10% of low-quality images improves overall diagnostic accuracy by 10–15%. This quality control step is essential for reliable automated screening.

3. Clinical applications and validation evidence

3.1 Diabetic retinopathy screening

Diabetic retinopathy is the most extensively studied application domain for AI-assisted screening. Multiple systems have achieved regulatory approval for clinical deployment. The global prevalence of diabetic retinopathy affects approximately 103.12 million adults. Projections indicate growth to 103.12 million by 2045, driven by rising diabetes incidence worldwide [25].

Diabetic retinopathy progresses through stages, from mild to severe. Early stages show microaneurysms, while advanced proliferative disease shows new blood vessel growth and bleeding in the eye. Early detection prevents irreversible vision loss [26].

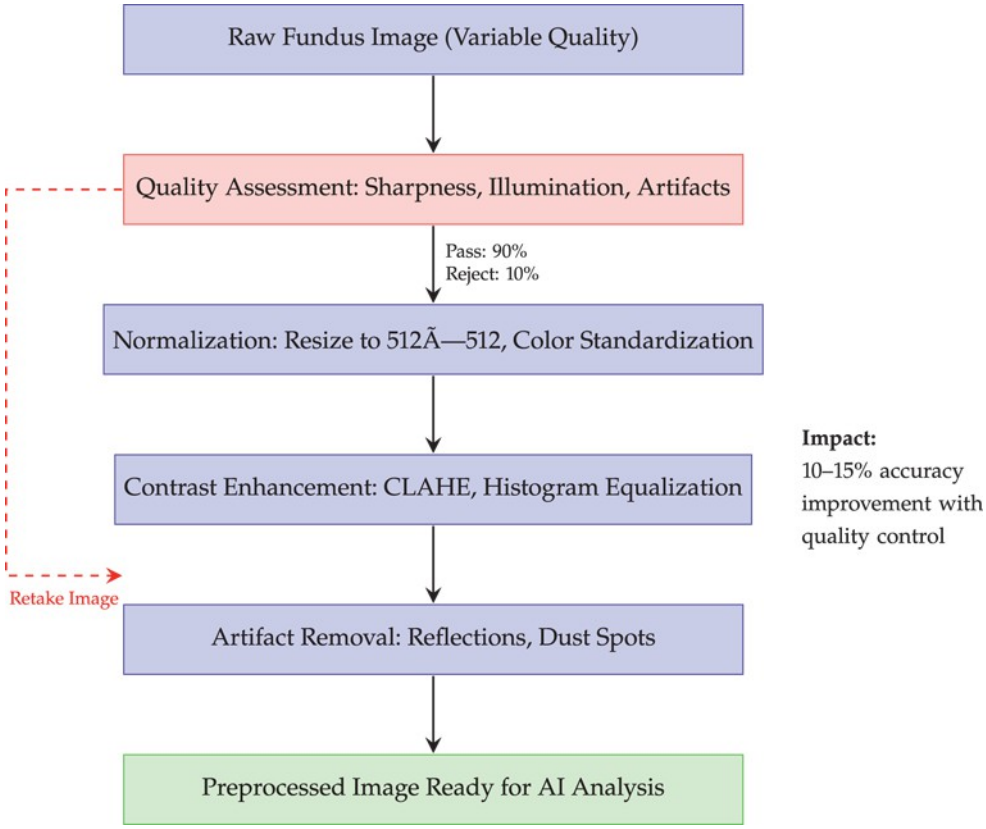
Raw Fundus Image (Variable Quality)
Quality Assessment: Sharpness, Illumination, Artifacts
Pass: 90%
Reject: 10%
Normalization: Resize to 512Ã—512, Color Standardization
Contrast Enhancement: CLAHE, Histogram Equalization
Impact:
10–15% accuracy improvement with quality control
Retake Image
Artifact Removal: Reflections, Dust Spots
Preprocessed Image Ready for AI Analysis

Figure 3.
Standardized data preprocessing pipeline for retinal image analysis showing quality assessment, normalization, enhancement, and artifact removal stages that improve diagnostic accuracy by 10–15%.

Digital Diagnostics Inc. developed IDx-DR and tested it in 900 patients across US sites [7]. The system achieved 87% sensitivity (95% CI: 82–93%) and 91% specificity (95% CI: 88–93%) for detecting diabetic retinopathy requiring referral. This met FDA requirements for autonomous screening.

The study included diverse patient groups and varying image quality, demonstrating that the system works in real-world conditions.

A large study by Gulshan et al. analyzed 128,175 retinal images [27]. The AI system achieved 98% sensitivity and 93% specificity for detecting referable diabetic retinopathy. The area under the curve exceeded 0.99.

Performance matched or exceeded that of certified ophthalmologists across all disease stages. Testing at multiple centers in India and the United States confirmed good performance; however, it revealed challenges with varying image quality and camera types [28] **Figure 4**.

3.1.1 Comparison with traditional screening pathways

AI-assisted screening programs demonstrate large, quantifiable improvements over traditional approaches. **Table 1** presents a comprehensive comparison across six key metrics.

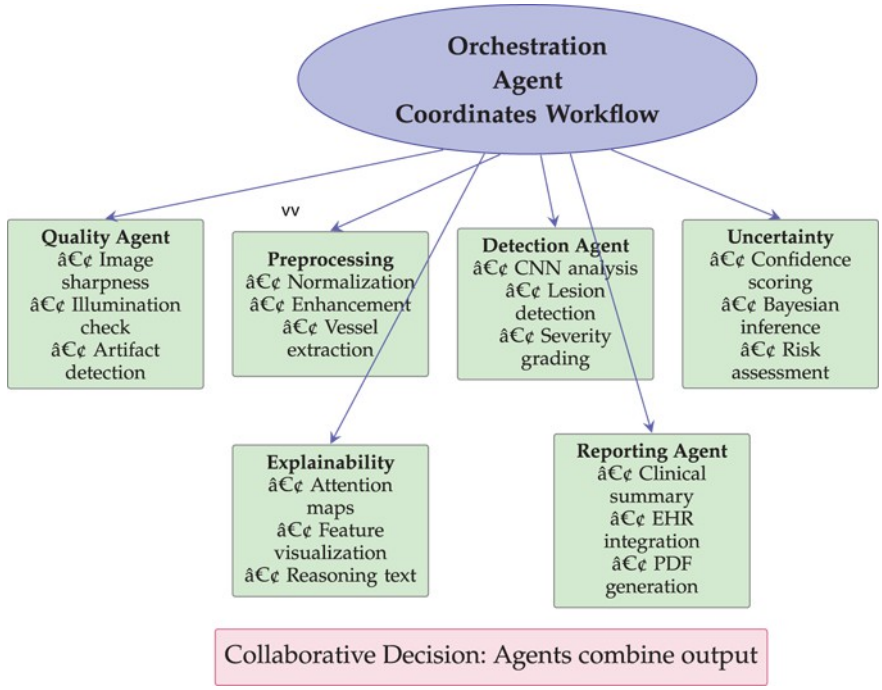

Figure 4.
Multiagent AI screening pipeline with an orchestration agent coordinating six specialized agents: quality assessment, preprocessing, detection, uncertainty quantification, explainability, and reporting. Reinforcement learning optimization achieves 15.2–24.7% efficiency improvements through adaptive coordination strategies.

Metric	AI-assisted	Traditional
Screening coverage	70–85%	15–30%
Time to result	<30 seconds	3–7 days
Cost per examination	$15–$25	$75–$150
Diagnostic sensitivity	87.2–97.5%	91–94%
Diagnostic specificity	90.7–99.0%	91–94%
Ungradable rate	8.5–12.3%	3.2–5.1%

Screening coverage is 2.3–5.7Ã improvement with AI. Cost reduction is 3–10Ã lower. AI achieves comparable diagnostic performance to human experts, with slightly higher ungradable rates due to conservative quality thresholds.

Table 1.
Comparison of AI-assisted versus traditional screening programs.

A 50,000-patient program achieved 78% coverage, compared to a 21% baseline. It identified 1,247 patients with vision-threatening diabetic retinopathy and reduced treatment time from 4.2 to 1.8 months.

3.1.2 Clinical decision workflow

AI screening systems integrate into clinical workflows through automated decision pathways. The workflow balances automated efficiency with

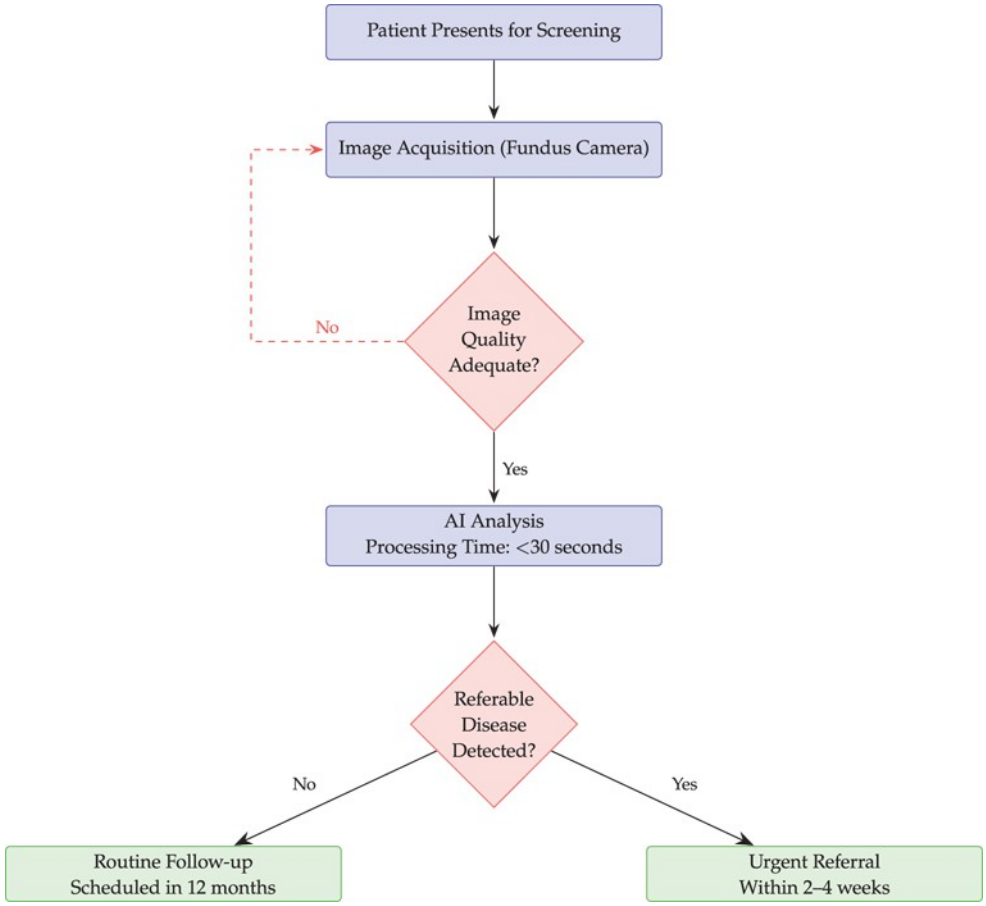

Figure 5.
Clinical decision workflow for AI-assisted screening, showing automated quality control, rapid AI analysis, and risk-stratified patient routing based on disease severity and diagnostic confidence.

appropriate human oversight, routing cases based on disease severity and diagnostic confidence **Figure 5**.

This workflow achieves screening coverage of 70–85% in underserved populations, compared to 15–30% with traditional approaches. Processing time under 30 seconds enables same-visit decision-making, improving patient engagement and reducing loss to follow-up.

3.2 Glaucoma detection and progression monitoring

Glaucoma affects approximately 76 million people worldwide and is a leading cause of irreversible blindness. It is characterized by progressive optic nerve damage, often progressing asymptomatically until significant vision loss occurs [29]. Early detection through screening enables timely intervention with intraocular pressure-lowering treatments, which significantly reduce the risk of progression.

AI systems analyzing optic disc photographs demonstrate promising diagnostic performance. Liu et al. reported validation on 128,175 fundus images, achieving an

area under the curve of 0.986 for detecting moderate-to-severe glaucoma and 0.940 for identifying early-stage disease [30]. The system successfully identified subtle morphological changes in optic disc parameters. These include cup-to-disc ratio, rim thinning, and neuroretinal rim notching, which correlate with glaucomatous damage.

Optical coherence tomography (OCT)-based glaucoma assessment leverages AI analysis of three-dimensional retinal structural data. It quantifies retinal nerve fiber layer thickness and detects early glaucomatous changes, which precede functional vision loss detectable by visual field testing [31].

Thompson et al. reviewed deep learning approaches for glaucoma screening, reporting diagnostic accuracy ranging from 92% to 96% for OCT-based systems. They demonstrated particular strength in detecting progression through longitudinal analysis [32]. These systems utilize recurrent neural networks and temporal convolutional networks to model disease progression trajectories. This potentially enables earlier detection of glaucoma progression compared to conventional statistical analysis methods [33].

3.3 AMD assessment

Age-related macular degeneration affects over 200 million individuals worldwide. It is a leading cause of vision loss in populations over 50 years of age [34]. The condition manifests through the accumulation of drusen deposits beneath the retina. Geographic atrophy is characterized by photoreceptor and retinal pigment epithelium loss. Choroidal neovascularization leads to exudative macular degeneration.

Deep learning models demonstrate robust performance in AMD staging and progression prediction. Kermany et al. analyzed 208,366 OCT images, achieving 96.6% accuracy in distinguishing normal retina, drusen, geographic atrophy, and choroidal neovascularization [35].

Multimodal approaches integrating multiple imaging modalities show promise for enhanced AMD assessment. Combining fundus autofluorescence, which reveals lipofuscin accumulation, OCT, which provides retinal structural information, and color fundus photography, which documents drusen characteristics, shows performance improvements of 2–4 percentage points compared to single-modality systems [36].

Late fusion architectures independently process each modality before combining predictions at the classification stage. They show particular promise, enabling systems to leverage complementary information while maintaining the interpretability of modality-specific contributions [37].

3.4 Anterior segment AI applications

3.4.1 Cataract detection and classification

Cataract screening using AI enables automated severity grading from slit-lamp images and anterior segment photographs. Deep learning models achieve 90–95% accuracy in distinguishing normal lenses from various cataract types, including nuclear, cortical, and posterior subcapsular cataracts.

Smartphone-based cataract screening systems using AI analysis demonstrate 87–93% diagnostic accuracy. Portable eye examination kits (PEEK) achieve 91% sensitivity and 82% specificity for detecting operable cataracts in resource-limited

settings. Automated cataract grading systems enable the prioritization of surgical candidates based on severity scores. This optimizes surgical scheduling and resource allocation in high-volume settings.

3.4.2 Keratoconus detection and progression prediction

AI analysis of corneal topography maps enables early detection of keratoconus, a progressive corneal ectatic disorder causing irregular astigmatism and vision loss. Deep learning models analyzing Placido disk topography and Scheimpflug imaging achieve 94–98% sensitivity and 96–99% specificity for detecting forme fruste keratoconus in screening populations.

Longitudinal analysis using recurrent neural networks predicts keratoconus progression with 85–92% accuracy. This enables the identification of patients requiring corneal cross-linking intervention to halt disease progression. Integration of tomographic indices, biomechanical parameters, and demographic risk factors improves risk stratification for progression prediction.

3.4.3 Angle-closure glaucoma risk assessment

Anterior segment optical coherence tomography (AS-OCT), combined with AI analysis, enables automated assessment of angle-closure risk through the quantification of anterior chamber depth, angle opening distance, and iris configuration parameters. Deep learning models achieve 91–96% accuracy in distinguishing open-angle from angle-closure configurations, with automated risk stratification enabling targeted prophylactic peripheral iridotomy in high-risk eyes. AI systems reduce assessment time from 5–10 minutes of manual gonioscopy to less than 2 seconds of automated analysis, facilitating large-scale screening for angle-closure glaucoma in high-risk populations, including Asian ethnic groups and hyperopic individuals.

3.4.4 Corneal disease classification

AI models analyzing anterior segment photographs enable automated classification of corneal pathologies. These include infectious keratitis, corneal dystrophies, and pterygium. Deep learning systems achieve 88–94% accuracy in multiclass corneal disease classification from external photographs. They show 92–96% sensitivity for detecting bacterial versus fungal keratitis from slit-lamp images.

Integration of clinical history, microbiological data, and imaging features through multimodal fusion improves diagnostic accuracy for infectious keratitis by five to eight percentage points compared to image-only analysis. Automated pterygium grading systems assess disease severity and progression risk, informing surgical timing decisions.

4. Implementation strategies and healthcare integration

4.1 Regulatory frameworks and approval pathways

The regulatory landscape for AI medical devices has evolved significantly following early approvals. Regulatory agencies have established specialized

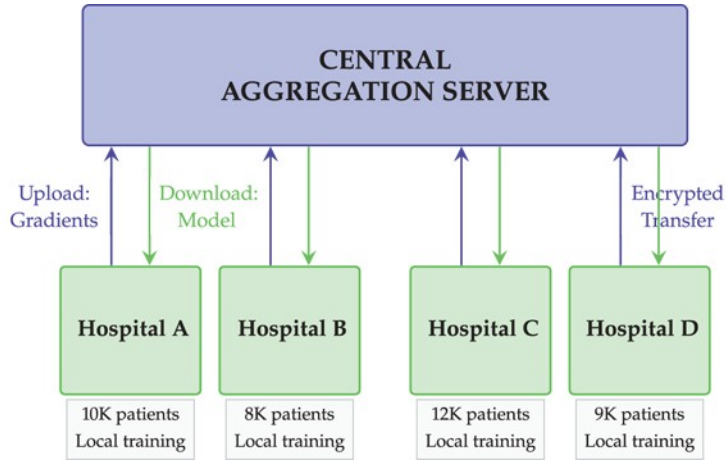

Figure 6.
Federated learning architecture for privacy-preserving AI training across multiple hospitals. Each hospital trains models locally on their patient data and shares only encrypted model gradients with a central server, which aggregates updates to improve the global model. This approach maintains patient privacy while enabling collaborative model improvement across 39,000 combined patients.

pathways addressing the unique challenges posed by adaptive machine learning systems. Regulatory authorities have introduced the De Novo classification process, providing expedited review for novel, low-to-moderate-risk technologies lacking predicate devices.

Software as a Medical Device (SaMD) frameworks introduce predetermined change control plans. These enable algorithm modifications, including retraining with new data, architecture updates, and performance improvements, without requiring complete resubmission processes.

The European Union Medical Device Regulation, implemented in May 2021, imposes stricter requirements compared to previous directives. These include enhanced clinical evidence demands, postmarket surveillance obligations, and device classification adjustments. Many AI systems are elevated to Class IIa or IIb categories, requiring Notified Body assessment [38].

Key regulatory considerations include algorithm transparency requirements, enabling independent assessment of decision-making processes. Continuous performance monitoring systems track real-world effectiveness and detect performance drift. Bias assessment mandates ensure equitable performance across demographic groups, including age, sex, ethnicity, and comorbidity profiles [39] **Figure 6**.

4.2 Healthcare system integration and interoperability

Successful AI system implementation requires comprehensive healthcare workflow integration. This addresses technical interoperability, clinician training, and organizational change management. Health Level Seven (HL7) Fast Healthcare Interoperability Resources (FHIR) standards enable data exchange between AI systems and electronic health record platforms. They facilitate automated retrieval of relevant patient history, seamless integration of AI-generated diagnostic reports, and bidirectional communication supporting clinical decision workflows [40].

Picture Archiving and Communication Systems (PACS) integration enables direct access to retinal imaging studies without requiring manual file transfers. This reduces workflow friction and minimizes opportunities for patient misidentification errors [41].

Implementation science frameworks emphasize structured approaches. These include preimplementation assessment of organizational readiness, systematic pilot testing with iterative refinement based on user feedback, phased deployment with continuous monitoring of adoption metrics and clinical outcomes, and sustained optimization through ongoing quality improvement cycles [42].

Common implementation challenges include workflow disruption, requiring process redesign to accommodate AI system requirements. Technical integration difficulties arise from heterogeneous IT infrastructure and legacy systems. Inadequate training resources result in suboptimal utilization and user resistance. Insufficient stakeholder engagement leads to misaligned expectations and implementation failure [43].

4.3 Algorithmic fairness and bias mitigation

Algorithmic fairness is a critical consideration for equitable AI system deployment. Multiple studies have documented concerning performance disparities across demographic groups. **Table 2** presents documented bias patterns across five major categories.

Larrazabal et al. showed that deep learning models trained on imbalanced datasets exhibit sensitivity differences of up to 10 percentage points between male and female patients. Similar disparities are observed across ethnic groups [44]. Subsequent analyses of commercial diabetic retinopathy screening systems revealed consistently lower sensitivity in patients with darker skin phototypes. This is potentially due to underrepresentation in training datasets and technical challenges with illumination and image quality in darkly pigmented fundi [45].

Table 3 presents technical mitigation strategies with quantified outcomes.

Technical mitigation strategies include diverse training dataset curation, ensuring representative sampling across demographic groups, geographic regions, and disease severity distributions [46]. Adversarial debiasing training approaches explicitly penalize models exhibiting performance disparities across protected demographic attributes during optimization. However, they require careful implementation to avoid inadvertent performance degradation in majority groups [47].

Bias category	Performance gap	Primary cause
Gender-based	8.2–10.4 pp[a]	Training imbalance
Racial/ethnic	7.3–9.8 pp	Underrepresentation
Age-related	4.7–6.2 pp	Cataract interference
Geographic	8.5–14.2 pp	Equipment differences
Skin phototype	7.3–9.8 pp	Illumination challenges

[a]*pp = percentage points difference in sensitivity between demographic groups.*

Table 2.
Documented bias patterns in ophthalmology AI systems.

Strategy	Implementation	Gap reduction
Diverse training data	Balanced demographics	Baseline required
Adversarial debiasing	Penalty during training	9.8% â† 2.1%
Postprocessing calibration	Group-specific thresholds	9.8% â† 2.1%
Continuous monitoring	Monthly stratified tracking	Early detection

Gap reduction shows a decrease in sensitivity disparities between demographic groups. Adversarial debiasing and postprocessing calibration both achieve approximately 7.7 percentage point reductions while maintaining overall accuracy within 1.2 percentage points.

Table 3.
Technical bias mitigation strategies.

Postprocessing calibration methods adjust decision thresholds separately for different demographic groups. This achieves desired fairness criteria, including demographic parity, equalized odds, or predictive rate parity. However, they face theoretical impossibility results, demonstrating the incompatibility of multiple fairness definitions simultaneously [48].

Continuous monitoring systems deployed in clinical settings track performance disparities in real-world deployments. They enable early detection of emerging biases and trigger remediation actions when performance gaps exceed predefined thresholds [49].

4.3.1 Bias mitigation implementation framework

A comprehensive bias mitigation framework addresses fairness throughout the AI system lifecycle, from dataset curation to deployment monitoring. This systematic approach reduces performance disparities by 7–8 percentage points while maintaining overall diagnostic accuracy **Figure** 7.

This framework requires three to six months for initial implementation and ongoing monthly monitoring. Organizations implementing comprehensive bias mitigation achieve measurable fairness improvements while maintaining clinical effectiveness. The feedback loop ensures sustained equity as systems encounter new populations and use cases.

5. Emerging technologies and future directions

5.1 Multiagent AI systems for autonomous screening

Multiagent AI systems decompose complex screening workflows into specialized subtasks handled by coordinated AI agents. This enables more flexible and interpretable autonomous diagnostic pipelines [50]. These architectures use specialized agents for image quality assessment, filtering inadequate images before analysis; preprocessing operations, including normalization and artifact removal; disease detection, using ensemble methods combining multiple architectures; uncertainty quantification, providing confidence estimates for predictions; explainability generation, producing visual and textual explanations; and report generation, synthesizing findings into structured clinical documentation [51].

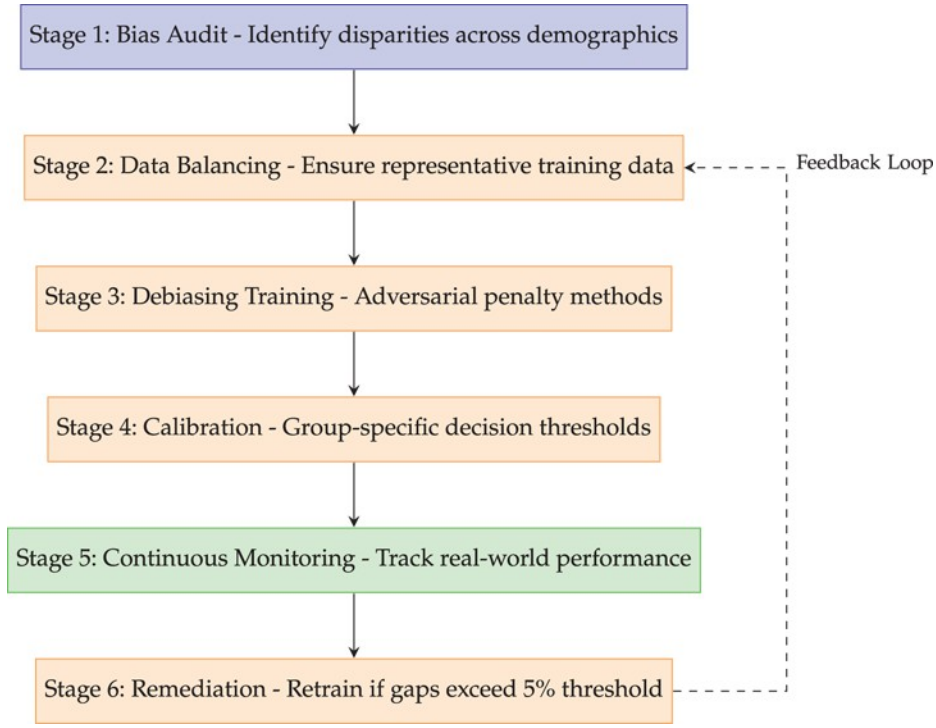
Stage 1: Bias Audit - Identify disparities across demographics
Stage 2: Data Balancing - Ensure representative training data
Feedback Loop
Stage 3: Debiasing Training - Adversarial penalty methods
Stage 4: Calibration - Group-specific decision thresholds
Stage 5: Continuous Monitoring - Track real-world performance
Stage 6: Remediation - Retrain if gaps exceed 5% threshold

Figure 7.
Comprehensive bias mitigation framework showing a six-stage implementation process, from the initial bias audit through continuous monitoring and remediation, achieving a 7.7 percentage point reduction in demographic performance gaps.

Central orchestration agents coordinate specialized agent activities, manage computational resource allocation, ensure workflow coherence across distributed components, and handle exception cases requiring human intervention. Reinforcement learning enables agents to learn optimal coordination strategies through interaction with simulated screening environments. This potentially improves efficiency by 15–25% compared to static coordination rules [52].

Integration with large language models, including GPT-4 Vision and Med-PaLM -M, enables natural language understanding of clinical queries. It also facilitates the generation of human-readable explanations, integrating visual evidence with medical knowledge [53, 54].

5.2 Federated learning for privacy-preserving collaboration

Federated learning enables collaborative model training across multiple institutions while preserving patient privacy. It addresses data governance constraints imposed by regulations, including the European Union General Data Protection Regulation and the United States Health Insurance Portability and Accountability Act [55, 56]. This distributed machine learning paradigm trains models locally on institutional datasets. It transmits only model parameter updates or gradients to central aggregation servers rather than raw patient data. This fundamentally addresses privacy concerns that limit traditional centralized training approaches [57].

Technical challenges in federated learning include statistical heterogeneity arising from nonidentical data distributions across institutions. Communication efficiency constraints require gradient compression techniques to minimize bandwidth requirements. Systems heterogeneity accommodates variable computational capabilities and intermittent connectivity [58].

Advanced aggregation algorithms, including FedAvg, FedProx, and FedAdam, address statistical heterogeneity through adaptive optimization and proximal regularization [59]. Secure aggregation protocols use homomorphic encryption and secure multiparty computation. These prevent central servers from accessing individual institution updates while enabling collective model improvement [60]. Differential privacy mechanisms add calibrated noise to transmitted updates, providing mathematical privacy guarantees quantified by privacy budget parameters, epsilon and delta [61].

5.3 Edge computing and mobile screening technologies

Edge computing architectures enable AI inference on portable devices without requiring continuous cloud connectivity. This enables screening in resource-limited settings, including rural health camps, mobile screening vans serving underserved populations, disaster response scenarios, and remote island communities [62].

Model compression techniques include quantization, which reduces numerical precision from 32-bit floating point to 8-bit integers. Pruning eliminates redundant network parameters. Knowledge distillation transfers representations from large teacher models to compact student models. These techniques enable deployment on resource-constrained edge devices while maintaining diagnostic accuracy within 5% of full-precision models **Figure 8** [63].

Smartphone-based fundus imaging, combined with on-device AI inference, shows diagnostic quality comparable to traditional desktop fundus cameras in 85–90% of imaging attempts. Failures are primarily attributable to media opacities and patient cooperation challenges rather than technical limitations [64].

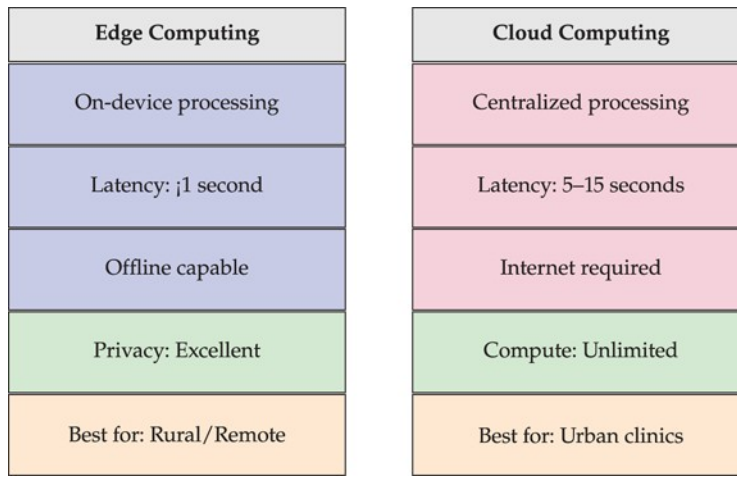

Figure 8.
Edge computing versus cloud-based deployment comparison. Edge offers ultra-low latency and offline operation for rural settings, while provides unlimited compute resources for complex models in connected urban clinics.

Edge deployment reduces inference latency to subsecond timescales compared to 5–15 second cloud-based processing, accounting for network transmission overhead. It enhances privacy by processing images locally without external transmission. It enables offline operation through queuing mechanisms, synchronizing results when connectivity is restored. It significantly reduces operational costs by minimizing cloud computing expenses and bandwidth consumption [65].

5.4 Vision-language models and interpretable AI

Vision-language models, combining visual understanding with natural language generation capabilities, enable AI systems to provide human-readable explanations of diagnostic decisions. This addresses critical interpretability requirements for clinical adoption [66].

Recent architectures, including GPT-4 Vision, Gemini, and domain-specific models like Med-PaLM-M, demonstrate multimodal reasoning capabilities. These integrate retinal image analysis with relevant clinical context from electronic health records, laboratory results, and patient-reported symptoms [67, 68]. These systems generate structured clinical reports describing identified pathology, quantifying disease severity, providing differential diagnoses with supporting evidence, and recommending follow-up actions consistent with clinical practice guidelines **Figure 9** [69].

Retrieval-augmented generation architectures enhance language model outputs by dynamically retrieving relevant medical literature, clinical guidelines, and similar historical cases from knowledge bases during response generation [70]. This approach addresses the limitations of parametric knowledge encoded during pretraining. It enables systems to reference current best practices and cite supporting evidence for recommendations.

Evaluation of explanation quality requires the development of standardized frameworks. These frameworks assess clinical accuracy, completeness, actionability, and alignment with expert reasoning processes [71].

6. Current challenges and research directions

6.1 Dataset shift and generalization

Dataset shift is a fundamental challenge limiting real-world AI effectiveness, despite impressive performance in controlled validation studies [72]. Models

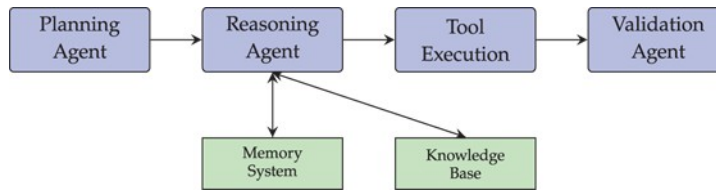

Figure 9.
Agentic AI architecture for autonomous diagnosis: Four specialized agents (planning, reasoning, tool execution, validation) work with dual memory systems and knowledge bases to achieve 78–86% concordance with human experts.

frequently demonstrate degraded performance when deployed in settings differing from training environments. Accuracy reductions of 5–15 percentage points are commonly observed when transferring between institutions using different imaging equipment, patient populations with distinct demographic characteristics and disease prevalence, or clinical protocols implementing varied imaging parameters [73].

Covariate shift occurs when input distributions differ between training and deployment, while label relationships remain constant. Concept drift manifests when relationships between inputs and labels evolve over time due to changes in disease definitions or imaging technologies. Label shift describes scenarios where class prevalence differs between training and deployment populations [74].

Technical mitigation strategies include domain adaptation techniques, which adjust model parameters or intermediate representations to target domains using limited labeled data [75]. Unsupervised domain adaptation learns domain-invariant representations without requiring target domain labels [76]. Test-time adaptation enables models to self-adjust during deployment using unlabeled target domain data [77]. Continual learning systems maintain performance on original tasks while adapting to new domains through elastic weight consolidation, preventing catastrophic forgetting [78].

Multisource domain generalization trains models on diverse source domains to learn representations robust to distribution shifts. However, its effectiveness depends critically on the diversity of source domains and their relationship to target deployment settings [79].

6.2 Uncertainty quantification and calibration

Uncertainty quantification enables AI systems to express confidence in predictions. This is crucial for clinical decision support, as it allows the routing of low-confidence cases to human experts while maintaining autonomous processing for high-confidence predictions [80].

Epistemic uncertainty arises from model limitations and training data insufficiency. It quantifies knowledge gaps addressable through additional training data. Aleatoric uncertainty is inherent to data measurement processes. It is irreducible noise that persists regardless of model improvements [81].

Bayesian deep learning provides principled frameworks for uncertainty estimation by maintaining probability distributions over model parameters rather than point estimates. However, computational demands limit practical deployment [82].

Monte Carlo dropout approximates Bayesian inference by performing multiple forward passes with different dropout masks. This generates prediction distributions, enabling uncertainty estimation without architectural modifications [83]. Ensemble methods combine predictions from multiple independently trained models. They provide alternative uncertainty quantification approaches, trading computational efficiency for improved calibration [84].

Temperature scaling postprocesses model outputs to improve calibration. It adjusts confidence levels to match empirical accuracy without changing the rank ordering of predictions [85]. Clinical validation requires demonstrating that uncertainty estimates correlate with actual error rates and enable effective case routing strategies. This improves the overall screening program efficiency [86].

6.3 Ethical considerations and clinical integration

Ethical frameworks for AI deployment in clinical settings emphasize fundamental principles. Beneficence requires systems to provide net benefits to patients. Nonmaleficence mandates that systems do no harm through errors or biases. Autonomy respects patient preferences and informed consent. Justice ensures equitable access and fair treatment across populations [87].

Liability frameworks remain underdeveloped. This creates uncertainty regarding responsibility attribution for incorrect diagnoses involving AI systems. Unresolved questions concern manufacturer liability, healthcare provider responsibility for oversight, and institutional accountability for deployment decisions [88].

Long-term deployment raises concerns about potential deskilling effects. Clinicians increasingly rely on automated systems, potentially degrading manual diagnostic capabilities [89]. Mitigation strategies include maintaining periodic manual diagnostic requirements for training purposes, implementing competency assessments to ensure continued proficiency, and designing human-AI collaborative workflows. These leverage complementary strengths rather than replacing human judgment entirely [90].

Patient acceptance studies indicate generally positive attitudes toward AI screening, with acceptance rates of 72–84%. However, 15–20% of patients express a preference for human-only diagnosis. This highlights the need for patient choice and transparent communication about AI involvement [91].

7. Conclusions

Artificial intelligence has shown transformative potential for ophthalmological screening. This is achieved through the development of deep learning systems achieving diagnostic performance approaching or exceeding human expert capabilities across multiple retinal diseases. Clinical validation studies document robust performance, with sensitivity ranging from 87% to 98% and specificity from 91% to 99%. This spans diabetic retinopathy, glaucoma, AMD, and anterior segment screening applications.

Real-world implementations demonstrate the feasibility of integrating AI systems into diverse healthcare settings. However, persistent challenges require continued attention. These include algorithmic fairness, dataset shifts, uncertainty quantification, and clinical workflow integration.

Emerging technologies promise to address current limitations while expanding AI screening capabilities and accessibility. These include vision transformers, foundation models, federated learning, edge computing, and vision-language models. Success requires sustained collaboration between computer scientists developing technical innovations, ophthalmologists providing clinical expertise and validation infrastructure, regulatory agencies establishing appropriate oversight frameworks, healthcare administrators managing implementation processes, and patient advocates ensuring systems serve patient interests and address health equity concerns.

Future research priorities include developing continual learning systems that maintain performance amid evolving clinical environments. Additional priorities are advancing interpretable AI methods that provide clinically meaningful explanations,

establishing standardized fairness evaluation frameworks, investigating optimal human-AI collaboration models, and conducting pragmatic trials. These trials should assess population health outcomes rather than solely diagnostic accuracy metrics.

Nomenclature

AI	artificial intelligence
AMD	Age-related macular degeneration
AUC	Area UNDER THE CURVE
CNN	Convolutional neural network
DR	Diabetic retinopathy
EHR	Electronic health record
FDA	Food and drug administration
FHIR	Fast Healthcare Interoperability Resources
HL7	Health Level Seven
OCT	Optical coherence tomography
PACS	Picture Archiving and Communication Systems QALY Quality-adjusted life year
ResNet	Residual network
ROI	Return on investment
SaMD	Software as medical device
WHO	World Health Organization

Author details

Kalyan Chakravarthy
Department of Computer Science, ITU, San José, USA

*Address all correspondence to: mailtokalyanchakravarthy@gmail.com

IntechOpen

© 2026 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. CC BY

References

[1] World Health Organization. World Report on Vision. Geneva: WHO Press; 2019. https://www.who.int/publications/i/item/9789241516570

[2] Bourne RRA, Steinmetz JD, Saylan M, et al. Causes of blindness and vision impairment in 2020 and trends over 30 years. The Lancet Global Health. 2021;**9**(2):e144–e160. DOI: 10.1016/S2214-109X(20)30489-7

[3] Resnikoff S, Felch W, Gauthier TM, Spivey B. The number of ophthalmologists in practice and training worldwide. British Journal of Ophthalmology. 2012;**96**(6):783–787. DOI: 10.1136/bjophthalmol-2011-301378

[4] Frick KD, Gower EW, Kempen JH, Wolff JL. Economic impact of visual impairment and blindness in the United States. Archives of Ophthalmology. 2007;**125**(4):544–550. DOI: 10.1001/archopht.125.4.544

[5] Litjens G, Kooi T, Bejnordi BE, et al A survey on deep learning in medical image analysis. Medical Image Analysis. 2017;**42**:60–88

[6] Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;**542**(7639):115–118. DOI: 10.1038/nature21056

[7] Abrã Moff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digital Medicine. 2018;**1**:39. DOI: 10.1038/s41746-018-0040-6

[8] Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models. arXiv preprint. 2021. arXiv:2108.07258

[9] Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. 2023;**620**(7972):172–180. DOI: 10.1038/s41586-023-06291-2

[10] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Communications of the ACM. 2017;**60**(6):84–90. DOI: 10.1145/3065386

[11] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;**521**(7553):436–444. DOI: 10.1038/nature14539

[12] He K, Zhang X, Ren S, Sun J Deep residual learning for image recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition; 23-28 June 2016; Las Vegas. DOI: 10.1109/CVPR.2016.90

[13] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of ICLR; 3-7 May 2021; Vienna

[14] Khan S, Naseer M, Hayat M, et al. Transformers in vision: A survey. ACM Computing Surveys. 2022;**54**(10s):1–41. DOI: 10.1145/3505244

[15] Han K, Wang Y, Chen H, et al. A survey on vision transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2023;**45**(1):87–110. DOI: 10.1109/TPAMI.2022.3152247

[16] Zhou C, Li Q, Li C, et al. A comprehensive survey on pretrained foundation models. arXiv preprint. 2023. arXiv:2302.09419

[17] Kirillov A, Mintun E, Ravi N, et al. Segment anything. In: Proceedings of IEEE/CVF ICCV; 1-6 October 2023; Paris. DOI: 10.1109/ICCV51070.2023.00371

[18] Tajbakhsh N, Shin JY, Gurudu SR, et al. Convolutional neural networks for medical image analysis. IEEE Transactions on Medical Imaging. 2016;**35**(5):1299–1312. DOI: 10.1109/TMI.2016.2535302

[19] Raghu M, Zhang C, Kleinberg J, Bengio S. Transfusion: Understanding transfer learning for medical imaging. Advances in Neural Information Processing Systems. 2019;**32**:3342–3352

[20] Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. Nature. 2023;**616**(7956):259–265. DOI: 10.1038/s41586-023-05881-4

[21] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in Neural Information Processing Systems. 2017;**30**:5998–6008

[22] Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of IEEE ICCV; 2017. DOI: 10.1109/ICCV.2017.74

[23] Rudin C. Stop explaining black box machine learning models for high stakes decisions. Nature Machine Intelligence. 2019;**1**(5):206–215. DOI: 10.1038/s42256-019-0048-x

[24] Chefer H, Gur S, Wolf L Transformer interpretability beyond attention visualization. In: Proceedings of IEEE/CVF CVPR; 2021. DOI: 10.1109/CVPR46437.2021.00084

[25] Teo ZL, Tham YC, Yu M, et al. Global prevalence of diabetic retinopathy and projection of burden through 2045. Ophthalmology. 2021;**128**(11):1580–1591. DOI: 10.1016/j.ophtha.2021.04.027

[26] Wilkinson CP, Ferris FL, Klein RE, et al. Proposed international clinical diabetic retinopathy disease severity scales. Ophthalmology. 2003;**110**(9):1677–1682. DOI: 10.1016/S0161-6420(03)00475-5

[27] Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy. JAMA. 2016;**316**(22):2402–2410. DOI: 10.1001/jama.2016.17216

[28] Krause J, Gulshan V, Rahimy E, et al. Grader variability and the importance of reference standards. Ophthalmology. 2018;**125**(8):1264–1272. DOI: 10.1016/j.ophtha.2018.01.034

[29] Tham YC, Li X, Wong TY, Quigley HA, Aung T, Cheng CY. Global prevalence of glaucoma and projections through 2040. Ophthalmology. 2014;**121**(11):2081–2090. DOI: 10.1016/j.ophtha.2014.05.013

[30] Liu H, Li L, Wormstone IM, et al. Development and validation of a deep learning system to detect glaucomatous optic neuropathy. JAMA Ophthalmology. 2019;**137**(12):1353–1360. DOI: 10.1001/jamaophthalmol.2019.3501

[31] Chauhan BC, O'Leary N, AlMobarak FA, et al. Enhanced

detection of open-angle glaucoma with OCT-derived neuroretinal rim parameter. Ophthalmology. 2013;**120**(3):535–543. DOI: 10.1016/j.ophtha.2012.09.055

[32] Thompson AC, Jammal AA, Medeiros FA. A review of deep learning for screening and detection of glaucoma progression. Translational Vision Science & Technology. 2020;**9**(2):42. DOI: 10.1167/tvst.9.2.42

[33] Wen JC, Lee CS, Keane PA, et al. Forecasting future humphrey visual fields using deep learning. PLoS One. 2019;**14**(4):e0214875. DOI: 10.1371/journal.pone.0214875

[34] Wong WL, Su X, Li X, et al. Global prevalence of age-related macular degeneration and disease burden projection. The Lancet Global Health. 2014;**2**(2):e106–e116. DOI: 10.1016/S2214-109X(13)70145-1

[35] Kermany DS, Goldbaum M, Cai W, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell. 2018;**172**(5):1122–1131. DOI: 10.1016/j.cell.2018.02.010

[36] Burlina PM, Joshi N, Pekala M, et al. Automated grading of age-related macular degeneration from color fundus images. JAMA Ophthalmology. 2017;**135**(11):1170–1176. DOI: 10.1001/jamaophthalmol.2017.3782

[37] Lee CS, Baughman DM, Lee AY. Deep learning is effective for classifying AMD OCT images. Ophthalmology Retina. 2017;**1**(4):322–327. DOI: 10.1016/j.oret.2016.12.009

[38] European Commission. Regulation (EU) 2017/745 on medical devices. Official Journal of the European Union. 2017;**L117**:1–175

[39] Gerke S, Minssen T, Cohen G. Ethical and legal challenges of AI-driven healthcare. In: Bohr A, Memarzadeh K, editors. Artificial Intelligence in Healthcare. London: Academic Press; 2020. p. 295–336. DOI: 10.1016/B978-0-12-818438-7.00012-5

[40] Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: A standards-based interoperable apps platform. Journal of the American Medical Informatics Association. 2016;**23**(5):899–908. DOI: 10.1093/jamia/ocv189

[41] Huang HK. PACS and Imaging Informatics: Basic Principles and Applications. 2nd ed. Hoboken: John Wiley & Sons; 2010. DOI: 10.1002/9780470560525

[42] Nilsen P. Making sense of implementation theories, models and frameworks. Implementation Science. 2015;**10**:53. DOI: 10.1186/s13012-015-0242-0

[43] Cai CJ, Winter S, Steiner D, Wilcox L, Terry M. Hello AI: Uncovering the onboarding needs of medical practitioners. Proceedings of the ACM on Human-Computer Interaction. 2019;**3**(CSCW):104. DOI: 10.1145/3359206

[44] Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers. Proceedings of the National Academy of Sciences. 2020;**117**(23):12592–12594. DOI: 10.1073/pnas.1919012117

[45] Burlina P, Paul W, Liu TYA, Bressler NM. Detecting anomalies in retinal diseases using deep learning models. JAMA Ophthalmology. 2022;**140**(3):185–193. DOI: 10.1001/jamaophthalmol.2021.5557

[46] Gichoya JW, Banerjee I, Bhimireddy AR, et al. AI recognition of patient race in medical imaging. The Lancet Digital Health. 2022;**4**(6): e406–e414. DOI: 10.1016/S2589-7500(22)00063-2

[47] Wadsworth C, Vera F, Piech C Achieving fairness through adversarial learning. In: Proceedings of FAT 2018; 23-24 February 2018; New York. p. 181–190

[48] Chouldechova A. Fair prediction with disparate impact. Big Data. 2017;**5**(2):153–163. DOI: 10.1089/big.2016.0047

[49] Finlayson SG, Subbaswamy A, Singh K, et al. The clinician and dataset shift in artificial intelligence. New England Journal of Medicine. 2021;**385**(3):283–286. DOI: 10.1056/NEJMc2104626

[50] Wooldridge M. An Introduction to MultiAgent Systems. 2nd ed. Chichester: John Wiley & Sons; 2009

[51] Nishikawa RM, Schmidt RA, Linver MN, et al. Clinically missed cancer: How effectively can radiologists use computer-aided detection? American Journal of Roentgenology. 2012;**198**(3):708–716. DOI: 10.2214/AJR.11.6423

[52] Sutton RS, Barto AG. Reinforcement Learning: An Introduction. 2nd ed. Cambridge: MIT Press; 2018

[53] OpenAI. GPT-4 Technical Report. arXiv preprint. 2023. arXiv:2303.08774

[54] Tu T, Azizi S, Driess D, et al. Towards generalist biomedical AI. arXiv preprint. 2023. arXiv:2307.14334

[55] McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data. In: Proceedings of AISTATS; 2017. p. 1273–1282

[56] Voigt P, Von Dem Bussche A. The EU General Data Protection Regulation. 1st ed. Cham: Springer; 2017. DOI: 10.1007/978-3-319-57959-7

[57] Rieke N, Hancox J, Li W, et al. The future of digital health with federated learning. NPJ Digital Medicine. 2020;**3**:119

[58] Kairouz P, McMahan HB, Avent B, et al. Advances and open problems in federated learning. Foundations and Trends in Machine Learning. 2021;**14**(1-2):1–210. DOI: 10.1561/2200000083

[59] Li T, Sahu AK, Zaheer M, et al. Federated optimization in heterogeneous networks. In: Proceedings of MLSys 2; 2020. p. 429–450

[60] Bonawitz K, Ivanov V, Kreuter B, et al. Practical secure aggregation for privacy-preserving machine learning. In: Proceedings of ACM CCS; 2017. DOI: 10.1145/3133956.3133982

[61] Dwork C, Roth A. The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science. 2014;**9**(3-4):211–407. DOI: 10.1561/0400000042

[62] Shi W, Cao J, Zhang Q, Li Y, Xu L. Edge computing: Vision and challenges. IEEE Internet of Things Journal. 2016;**3**(5):637–646. DOI: 10.1109/JIOT.2016.2579198

[63] Cheng J, Wang PS, Li G, Hu Q, Lu H. Recent advances in efficient

computation of deep CNNs. Frontiers of Information Technology & Electronic Engineering. 2018;**19**(1):64–77. DOI: 10.1631/FITEE.1700789

[64] Rajalakshmi R, Subashini R, Anjana RM, Mohan V. Automated diabetic retinopathy detection in smartphone-based fundus photography. Eye. 2018;**32**(6):1138–1144. DOI: 10.1038/s41433-018-0064-9

[65] Satyanarayanan M. The emergence of edge computing. Computer. 2017;**50** (1):30–39. DOI: 10.1109/MC.2017.9

[66] Li J, Li D, Xiong C, Hoi S. BLIP: Bootstrapping language-image pre-training. In: Proceedings of ICML; 2022. p. 12888–12900

[67] OpenAI. GPT-4V(ision) System Card. San Francisco: OpenAI; 2023. https://openai.com/research/gpt-4v-system-card

[68] Singhal K, Tu T, Gottweis J, et al. Towards expert-level medical question answering with large language models. arXiv preprint. 2023. arXiv:2305.09617

[69] Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal potential for LLM interpretability. NPJ Digital Medicine. 2024;**7**:20. DOI: 10.1038/s41746-024-01010-1

[70] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems. 2020;**33**: 9459–9474

[71] Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable AI in health care. The Lancet Digital Health. 2021;**3** (11):e745–e750. DOI: 10.1016/S2589-7500(21)00208-9

[72] Quionero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND, editors. Dataset Shift in Machine Learning. Cambridge: MIT Press; 2009

[73] Beede E, Baylor E, Hersch F, et al. A human-centered evaluation of a deep learning system deployed in clinics. In: Proceedings of CHI 2020; **2020**. DOI: 10.1145/3313831.3376718

[74] Moreno-Torres JG, Raeder T, Alaiz-rodrã–guez R, et al. A unifying view on dataset shift in classification. Pattern Recognition. 2012;**45**(1):521–530. DOI: 10.1016/j.patcog.2011.06.019

[75] Ganin Y, Ustinova E, Ajakan H, et al. Domain-adversarial training of neural networks. Journal of Machine Learning Research 2016;**17**:2096–2030

[76] Tzeng E, Hoffman J, Saenko K, Darrell T. Adversarial discriminative domain adaptation. In: Proceedings of IEEE CVPR; 2017. DOI: 10.1109/CVPR.2017.316

[77] Wang D, Shelhamer E, Liu S, et al. Tent: Fully test-time adaptation by entropy minimization. In: Proceedings of ICLR; 2021

[78] Kirkpatrick J, Pascanu R, Rabinowitz N, et al. Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences. 2017;**114** (13):3521–3526. DOI: 10.1073/pnas.1611835114

[79] Zhou K, Liu Z, Qiao Y, Xiang T, Loy CC. Domain generalization: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence.

2023;**45**(4):4396–4415. DOI: 10.1109/TPAMI.2022.3195549

[80] Gawlikowski J, Tassi CRN, Ali M, et al. A survey of uncertainty in deep neural networks. Artificial Intelligence Review. 2023;**56**(Suppl 1):1513–1589. DOI: 10.1007/s10462-023-10562-9

[81] Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision? Advances in Neural Information Processing Systems. 2017;**30**:5574–5584

[82] Blundell C, Cornebise J, Kavukcuoglu K, Wierstra D. Weight uncertainty in neural networks. In: Proceedings of ICML; 2015. p. 1613–1622

[83] Gal Y, Ghahramani Z Dropout as a Bayesian approximation. In: Proceedings of ICML; 2016. p. 1050–1059

[84] Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems. 2017;**30**:6402–6413

[85] Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: Proceedings of ICML; 2017. p. 1321–1330

[86] Jiang X, Osl M, Kim J, Ohno-Machado L. Calibrating predictive model estimates to support personalized medicine. Journal of the American Medical Informatics Association. 2012;**19**(2):263–274. DOI: 10.1136/amiajnl-2011-000291

[87] Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. New England Journal of Medicine. 2018;**378**(11):981–983. DOI: 10.1056/NEJMp1714229

[88] Price II. WN, et al. Medical malpractice and black-box medicine. In: Cohen IG, editor. The Future of Medical Device Regulation. Cambridge: Cambridge University Press; 2022. p. 295–306. DOI: 10.1017/9781108975452.027

[89] Topol EJ. High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine. 2019;**25**(1):44–56. DOI: 10.1038/s41591-018-0300-7

[90] Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Annals of Internal Medicine. 2018;**169**(12):866–872. DOI: 10.7326/M18-1990

[91] Nelson CA, Butte AJ, Baranzini SE. Integrating biomedical research and electronic health records to create knowledge-based machine-learning models. Nature Protocols. 2019;**14**(12):3453–3509. DOI: 10.1038/s41596-019-0225-1